Number of Pages:	31
Number of Words (narrative, acknowledgments):	5897
Number of References:	8
Number of Tables:	6
Number of Figures:	5

EFFECTIVENESS OF AN EXPERT SYSTEM FOR ASTRONAUT ASSISTANCE ON A SLEEP EXPERIMENT

Gianluca Callini[1], S.M., B.M.E. luca@mit.edu (1)

Susanne M. Essig, M.S., B.S.A.E. smessig@mit.edu (1)

Dennis M. Heher, M.S., B.A. heher@ptolemy-ethernet.arc.nasa.gov (2)

Laurence R. Young, Sc.D. lry@space.mit.edu (1)

(1) Man-Vehicle Laboratory, Massachusetts Institute of Technology

77 Massachusetts Institute of Technology, Room 37-219

Cambridge, MA 02139

Tel: (617) 253-7805

Fax: (617) 258-8111

(2) Caelum Research Corporation - NASA Ames Research Center

Mail Stop 269-2 (Building 269 Room 235)

Moffett Field, CA 94035

Tel: (650) 604-1084

Fax: (650) 604-3594

Address Manuscript Correspondence to Laurence R. Young, Sc.D.

Running Head: Evaluation of an Expert System

[1] Principal author, MIT Man-Vehicle Laboratory.

ABSTRACT

Background: Principal Investigator-in-a-Box ([PI]) is an expert system designed to train and assist astronauts with the performance of an experiment outside their field of expertise, particularly when contact with the Principal Investigators on the ground is limited or impossible. In the current case, [PI] was designed to assist with the calibration and troubleshooting procedures of the Neurolab Sleep and Respiration Experiment. [PI] displays physiological signals in real time during the pre-sleep instrumentation period, alerts the astronauts when a poor signal quality is detected, and displays steps to improve quality. **Methods:** Two studies are presented in this paper. In the first study twelve subjects monitored a set of prerecorded physiological signals and attempted to identify any signal artifacts appearing on the computer screen. Every subject performed the experiment twice, once with the assistance of [PI] and once without. The second part of this study focuses on the post-flight analysis of the data gathered from the Neurolab Mission. After replaying the physiological signals on the ground, the frequency of correct [PI] alerts and false alarms (i.e., incorrect diagnoses by the expert system) was determined in order to assess the robustness and accuracy of the rules. **Conclusions:** Results of the ground study indicated a beneficial effect of [PI] and training in reducing anomaly detection time and the number of undetected anomalies. For the in-flight performance, excluding the saturated signals, the expert system had an 84% detection accuracy, and the questionnaires filled out by the astronauts showed positive crew reactions to the expert system.

Index Terms: expert system, sleep, space, artificial intelligence.

INTRODUCTION

Background

In order to assure the effective performance of a physiological experiment in space, it is essential to provide the astronaut operator with adequate training and in-flight assistance, either by direct contact with the Principal Investigator or by use of some other means**.** During the course of a single mission, astronauts typically conduct several experiments outside their field of expertise. Errors in experimental procedure can cause inadequate data collection or even complete data loss. Since it is impractical to allow each Principal Investigator to fly into space with his or her experiment, or even to have ready accessibility to the astronaut and the experiment, Principal Investigator-in-a-Box (abbreviated [PI]) was created as an onboard decision aid for astronauts. This artificial intelligence computer system is designed to carry some of the Principal Investigator's knowledge into space for real-time access during the conduction of an experiment.

Many factors influence astronaut performance during a mission. Stress, fatigue, sleep loss (2, 5), delay between training and the actual performance of the experiment, and possible additional stress associated with long-duration space flights are only a few of these. It is common to find in-flight performance errors for procedures that were performed flawlessly during ground simulations. As long-duration space flights become more common, including those planned for the International Space Station, the related problems are certain to increase, making the role of such expert systems even more important in reducing the likelihood of astronaut error.

Although crew training schedules generally cover experiment normal operations and some malfunction procedures, there is rarely time to train astronauts for the adept and

innovative decision making unexpected problems often require. Many troubleshooting issues have been handled through direct or indirect communication with the PI on the ground, including exchange of data and graphics. However, this solution is not likely to be generally available for the Space Station. Open discussion of problems on public air-to-ground loops is inhibited. Communication restrictions and the often-limited accessibility of Principal Investigators speak to the importance of an AI countermeasure for use in onboard experiments.

Most expert systems use the same heuristics ("rules of thumb") and strategies that human experts use when troubleshooting or solving a problem. These *rule-based* systems mimic a level of human behavior in order to recognize patterns and activate their algorithms to execute the appropriate response. The most popular method for encoding such algorithms is by expressing the knowledge with a series of "if-then" rules.

Rule-based systems have many advantages for encoding knowledge. New rules can be added without disturbing the rest of the system, which enables uncomplicated modification of straight-line code. Rules are a good way to model the strong data-driven nature of intelligent action: as new data arrives, the behavior of the system changes. Rules achieve generality because their conditions and conclusions are expressed in a pattern-matching language where a single pattern may appropriately match a large variety of actual data. These systems are very efficient because well known indexing methods optimize the checking of rule conditions: only those rules whose conditions are relevant to newly asserted facts are triggered.

In areas where the expert task is data-driven (i.e., the system is guided by successive data or observations), a technique called "forward chaining" is employed. In

such systems, available data is continually matched against the “if” part (condition) of each rule, and when all conditions in the “if” part match the data, the corresponding conclusions in the “then” part are applied. New facts, whether arising from new observations, data input, or assertions by other rules, participate in the matching process and can lead, in turn, to the operation of additional rules. The opposite approach, called “backward chaining,” is goal-driven, and the goal as found in the rules’ conclusions is first checked to see what constraints need to be satisfied.

Expert systems have been used in a variety of applications such as planning and scheduling, diagnosis and troubleshooting of devices, decision-making, monitoring and control, and medical support (6). Expert systems can also be designed to implement applications simultaneously. Expert systems capable of analyzing incoming real-time data from a system with the purpose of noticing and reporting anomalies, as well as predicting trends, has been implemented in various fields. Despite the obvious advantages of such a system, real-time expert systems are not as common or widespread as their non-real-time counterparts.

Expert systems applications in space are relatively rare, although the Space Program has seen a few attempts at using expert systems in the engineering field to assist astronauts and ground operators with the performance of a mission (3, 4, 7). As on the ground, however, real-time space-related expert systems are virtually unprecedented. In fact, the earliest use of a real-time expert system for control of experiments in the field of life sciences is the direct predecessor of the expert system presented in this paper (1, 8).

Although the formal evaluation of a real-time space-related medical expert system such as [PI] is unprecedented, most of the methods used by our predecessors

apply. Evaluation of a clinical/medical system takes place over several iterative steps during the development cycle. In fact, the evaluation often continues even after the system has been publicly marketed or distributed.

Rationale for the Study

As discussed earlier, expert systems can make great contributions toward helping astronauts meet the special demands of conducting experiments during long-term space flight. Evaluation of a system designed to assist with medical diagnostics demonstrates the benefits an onboard expert system could offer to the life sciences community. Our first study on such a system was divided into two parts: a ground-based experiment that used student subjects as "astronaut surrogates" and a later experiment using the actual data gathered from the four science crew members on a Space Shuttle mission. The evaluation of [PI] for both the ground and flight studies focused on the speed and reliability of the human-computer system in the detection and identification of anomalies in the physiological signals monitored by [PI].

The first version of [PI], also known as the Astronaut Science Advisor (ASA), is the first documented attempt to use a biomedical diagnostic expert system on a space mission (1, 8). [PI] was used to assist astronauts in the performance of the "Rotating Dome" visual-vestibular interaction experiment on the STS-58 Space Life Sciences 2 (SLS-2) Space Shuttle mission in 1993. This first version of [PI] provided data collection capabilities, as well as protocol assistance, scheduling, and protocol modification suggestions. An additional feature consisted of an "interesting data" filter, designed to perform quick-look data analysis and report any unexpected findings to the astronauts

during the experiment. Although crew feedback on this demonstration was positive, no data was taken concerning the performance of [PI] or the correctness of the advisories that it issued.

Extending the successful implementation of the ASA with the Rotating Dome experiment, MIT and NASA Ames Research Center collaborated on the development of a new version of [PI] in conjunction with the "Sleep, Respiration and Melatonin in Microgravity" experiment (commonly referred to as the Sleep and Respiration experiment), led by Dr. Charles Czeisler of Brigham and Women's Hospital (BWH, Boston, MA) and Dr. John West of the University of California, San Diego. The experiment flew aboard the STS-90 (Neurolab) Space Shuttle mission in April-May 1998.

This new version of [PI], designed as an integral part of the experiment, assisted the Neurolab astronauts with the calibration and troubleshooting of the instrumentation (described in detail later) during the pre-sleep period of the experiment, when mission rules preclude investigator ground-to-air contact with the crew. Because the crucial experiment setup and calibration for the extended period of sleep monitoring is performed during this no-contact phase, the crew is necessarily isolated from the true science experts. During this phase of the experiment, [PI]'s role is to display the subjects' physiological signals, identify anomalous signals, and suggest corrective procedures when necessary.

The [PI] anomaly-identification process is achieved in several steps. First, the system calculates several standard statistical quantities (mean, variance, standard deviation) for every signal type. These values are then transferred to the reasoning

engine, containing specific rules for every signal. The system first checks for the presence of a reasonable physiological signal. If the calculated values do not meet the acceptable ranges as specified by the Sleep Team experts and incorporated in the rules, the system displays a red light for the corresponding signal and indicates the appropriate troubleshooting procedure. Subsequently, the system checks all signals for quality, using a different set of rules. The [PI] graphic user interface available to the astronauts is shown in **Fig. 1** below, with the electrophysiological and cardiorespiratory signals displayed on the screen.

[FIGURE 1 – HERE]

The Sleep experiment, accompanied by [PI], also flew aboard the STS-95 mission in October-November 1998 to study the effects on sleep of space flight and aging.

The Neurolab Sleep Experiment

A brief overview of the Neurolab Sleep Experiment is offered here as a means to understanding and appreciating [PI]'s function within the Neurolab and STS-95 experiments, as well as the results presented later. There are two halves of the Sleep and Respiration Experiment. The first half, devised by Dr. Czeisler, is to study the effect of melatonin on sleep in weightlessness. Eight electrophysiological signals are monitored and recorded to assess the characteristics of the astronauts' sleep. Four electroencephalogram (EEG) signals (brain waves), two electromyogram (EMG) signals

(muscle activity), and two electro-oculogram (EOG) signals (eye movements) are recorded. The second half of the experiment, developed by Dr. West, studies the effects of microgravity on respiration. This portion requires recording of a series of cardiorespiratory signals: electrocardiogram (EKG) or heart rate, blood oxygen saturation level (SaO_2), abdominal and ribcage expansions, nasal airflow and the presence of snoring sounds. Another cardiorespiratory signal recorded is referred to as "pulse-wave" or "PWave." Note that this "PWave" is totally unrelated to the section of the electrocardiogram wave corresponding to the atrial depolarization and contraction (also referred to as "PWave." The pulse-wave is measured with the same device used for the oxygen saturation reading. It appears as a rhythmic, sawtooth-shaped wave representing the sensor's measurement of the subject's pulsatile blood flow. This waveform is the basis of the oxygen saturation measurement. The amplitude and clarity of the signal indicate the quality of the sensor's reading. A small or noisy pulse-wave indicates something wrong with the sensor (e.g., placement), which may not be apparent from the SaO_2 signal. For this reason, the pulse-wave is used more as an indicator of the quality of the oxygen saturation signal, than as a direct measurement of a specific physiological function.

The astronauts work in teams of two to apply this instrumentation to each other during the pre-sleep period. The hardware consists of the following items:

- Electrode Net (e•Net): an elastic web-like cap containing 13 electrode sockets to record the EEG, EMG and EOG signals (Physiometrix, Inc., North Billerica, Massachusetts, U.S.A.)

- "Respiratory Inductance Plethysmography" (RIP) Suit: a Lycra tank top and shorts containing instrumentation to record abdominal and chest expansions (Blackbottom, Inc. California, U.S.A.);
- "Borg Harness": a bundle of electronic connections and cables for the RIP suit plus instrumentation to measure the nasal airflow via a nasal thermistor, EKG, the presence of snoring sounds via a microphone, and blood oxygen saturation level and pulse-wave signals via a pulse oximeter worn on a finger (manufactured at the Physiology/NASA Laboratory, The University of California, San Diego, U.S.A.);
- Digital Sleep Recorder (DSR): a device that converts the raw analog signals from the various electrodes and instrumentation to digital signals, which are then recorded onto a PCMCIA FlashRAM card (Copyright 1996 Vitaport EDV System GmbH. Distributed by TEMEC instruments BV, The Netherlands).

The flight computer on which [PI] was installed is an IBM ThinkPad 755C laptop equipped with an Intel 486-75 MHz processor and 20 MB of RAM. This hardware constraint, imposed by the requirement to use NASA flight hardware, was apparent during the development of [PI], which was actually coded on a much faster Intel Pentium-based computer. Some of the rules used by the expert system may appear somewhat simplistic. However it should be kept in mind that more complicated rules on a 486-based computer would have significantly slowed down the computation and prevented the system from displaying signals in real time.

During the pre-sleep calibration period, the [PI] laptop interfaces with the DSR via an RS-232 serial optical cable. A schematic diagram depicting the manner in which [PI] is connected to the rest of the flight hardware is shown below (**Fig. 2**).

[FIGURE 2 – HERE]

Equipment similar to the Neurolab Sleep hardware was used in the ground-based experiment with astronaut surrogates.

METHODS

Ground-Based Evaluation Goals[2]

The pilot study was performed to acquire preliminary results on the efficacy of [PI]. The experimental protocol was reviewed and approved by the MIT Committee on the Use of Humans as Experimental Subjects (COUHES), and informed consent was obtained from all the subjects involved in the experiment. Student subjects with minimal experiment training were used to test the hypothesis that an expert system such as [PI] would successfully assist users in the performance of a life sciences experiment outside their field of expertise. The results with [PI] were compared to a control condition that included training but no assistance from the expert system. A secondary goal was to identify specific aspects of [PI] that influenced subjects' performance during the experiment.

Subjects

Twelve subjects, six male and six female, took part in this experiment. The

[2] Large portions of the following *Description*, *Results*, and *Discussion* sections on the ground-based study are extracted from Gianluca Callini's Unpublished Master's thesis "Assessment of an Expert System for

subjects were all graduate students in the Department of Aeronautics and Astronautics at MIT. The mean age of the subjects was 25 years; only one subject was older than 30 years.

Protocol

The day before beginning experimental activities, the subjects attended a 1.5-hour training lecture. The training introduced the subjects to the identification of electrophysiological sleep data, including the detection of signal anomalies created by improper instrumentation setup or hardware malfunctions. The subjects were also introduced to [PI] and its diagnostic capabilities. A live demonstration was given to the subjects by having [PI] play a data file. The experiment was fully described and a short quiz was administered at the end of the session to assess the effectiveness of subject training. Most subjects received perfect scores. Although the training period was considerably less than the total time the Neurolab astronauts trained on the Sleep Experiment, it was comparable to the amount of time each crew member spent training on [PI].

Experiment Design

The subjects in the evaluation were divided into two groups of six, which began the experiment with and without [PI] respectively. They were each asked to monitor a set of pre-recorded electrophysiological signals and to detect and identify each signal artifact displayed on the screen, just as they would if they were actually performing the Neurolab

experiment by wearing the sleep instrumentation. Due to scheduling and time constraints, the groups were not balanced by gender. The first group (group A), composed of four males and two females, received [PI] assistance only on the first day. The second group (group B) received [PI] assistance only on the second day, and was composed of two males and four females. Acting as his or her own control, every subject performed the experiment with and without the help of [PI]'s diagnostic capabilities on two consecutive days. The groups performed the tests in a crossover fashion, as represented in **Table I**:

[TABLE 1 – HERE]

The subjects were provided with a reference manual containing a synopsis of the training session, as well as a list of the electrophysiological signals displayed on the screen and the anomalies detected by [PI]. After briefly reviewing the material covered in the training session, the subjects were instructed to start the test session, which lasted about twenty minutes. All twelve subjects completed the experiment, and no software or hardware failures were experienced.

The data file the subjects were asked to monitor consisted of real data recorded at the NASA Johnson Space Center during one of the Neurolab crew members' training sessions. The data file contains a total of 59 anomalies for the electrophysiological signals. At least one anomaly appeared on every electrophysiological signal type displayed on the screen. The duration of the anomalies varied, and there were several periods of time when all the signals displayed on the screen were nominal. Although the same file was used for all the tests on both days, there were no indications that the

subjects acquired enough familiarity with the random appearance of signal artifacts to influence their performances on the second experimental day. [PI] recorded every anomaly onset time and the corresponding subject reaction times.

Flight Performance Background

The Sleep and Respiration Experiment was performed on two separate four-day periods during the 16-day Neurolab mission in 1998. The four science astronauts on the mission were subjects for the sleep experiment, and therefore all used [PI]. Post-flight questionnaires were distributed to these four astronauts as an additional way to assess [PI]'s performance and crew interaction during the actual mission. The data set consisted of the first 15 minutes (pre-sleep) of the sleep signals recorded during each sleep session. Using [PI] on the ground, it was possible to replay these signals post-flight, record the anomaly onset times and, for each case, judge whether [PI]'s heuristics worked correctly, or if false alarms were generated. It was not feasible to check for missed detections (false negatives).

RESULTS

Ground-Based Study Results

An analysis of variance (ANOVA) was performed to determine the influence and significance of several factors on various aspects of subject performance. The average reaction times for the subjects to detect an anomaly, as well as the number of undetected anomalies for both groups, are plotted in **Fig. 3.** Group A ([PI] assistance on Day 1 only) results are shown in **Fig. 3** (a) and (c), while group B results ([PI] assistance on Day 2

only) are shown in **Fig. 3** (b) and (d). The dotted lines indicate results obtained by the subjects with the assistance of [PI], while solid lines indicate those obtained when [PI] was inactive.

[FIGURE 3 – HERE]

Members of group B performed the experiment without the assistance of [PI] on the first day and with the assistance of [PI] on the second. Most of these subjects showed a significant improvement in response time the second day, when [PI] was activated. The average response time for Group B decreased by nearly half on day two with [PI] assistance (**Fig. 3** (b)). Group A, however, which received assistance from [PI] diagnostics on day one, did not show a significant difference in average response time on day two, when [PI] assistance was no longer given (**Fig. 3** (a)). The average response time decreased only by a minimum amount on day two (without [PI] assistance). **Table II** below summarizes the statistical analysis for the average overall reaction time. The values reported in this table, as well as the following one have the following meaning:

- "n" is the number of cases used for every type of measure studied. The maximum number for this value is 59, since the data files recorded contained 59 anomalies;
- "t" is the pool variance obtained from the t-test and it indicates the significance of a given effect on a particular measurement;
- "F" is the F-ratio, which is the ratio of the mean square of each effect or cross-effect to the mean square for error;

- “p” is the p-value, which is the probability of exceeding the F-ratio and indicates the significance of a given effect on a measurement. An effect is defined as significant if the p value is less than 0.05;
- “Mean Effect” is the mean value of the effects on subject performance; a positive effect for a given condition indicates that the [PI] assistance or day effect decreased the reaction time or the number of undetected anomalies.

[TABLE 2 – HERE]

Statistically, the only effect on the average reaction time was the combination of [PI] assistance and day, suggesting that the subjects were able to detect signal anomalies about 10 seconds faster with [PI] on the second day, due to a positive influence of training. This would indicate a training effect on both the usage of [PI] and the monitoring of sleep signals. This is reflected in **Fig. 3** (b) where the subjects who used [PI] on the second day performed much better than the group that did not have [PI] available on the second day.

The number of undetected anomalies per subject per day was then analyzed to observe the direct effects and interactions of day and [PI] assistance. The number of undetected anomalies significantly decreased in group B when [PI] was active on the second day, as seen clearly in **Fig. 3** (d). For the subjects of group A, the number of undetected anomalies was also generally lower when [PI] was active (**Fig. 3** (c)). **Table III** below shows the effects on the number of undetected anomalies:

[TABLE 3 – HERE]

Note that both the effect of [PI] and the interaction of [PI] assistance and day were significant in decreasing the number of undetected anomalies. This also suggests an effect of training.

Flight Performance Results From Data File Analysis

A total of 16 sleep signal recordings were obtained from the Neurolab mission (one per subject per instrumentation session). The files were replayed on the ground to record all the signal artifacts encountered and to assess the accuracy of the [PI] diagnostics. The number of false alarms for each sleep session is tabulated below (**Table IV**).

[TABLE 4 – HERE]

For this analysis, the false alarms were determined by two trained data analysts (MIT graduate students and authors Gianluca Callini and Susanne Essig, who gained experience on signal monitoring by working in conjunction with BWH). They replayed the files and examined the data to judge [PI]'s diagnoses for each anomaly. False alarms were defined as cases in which [PI] would alert the astronauts to a poor quality signal when the signal display otherwise showed a good quality signal. In a number of cases, [PI] would activate a red state light for a simple signal saturation, since its rules were not necessarily coded to take that effect into account. Throughout the studies performed with [PI], saturation was defined as the condition in which the value of a signal exceeded the pre-established display ranges, and therefore appeared as a flat line at the very top or the very bottom of its display range; although it appeared flat, the signal was in fact still

present, though not visible. Because the signal was outside the ranges coded in the reasoning engine's rules, [PI] responded to saturation with a red light. These cases were tabulated separately from the false alarms, as the table shows. The astronauts, however, were trained in signal monitoring to varying degrees, and were expected to successfully distinguish a saturation signal, which would normally correct itself, from an actual alert requiring troubleshooting. The percentages of valid identifications of poor quality signals were calculated in two different ways, with and without accounting for the saturated signals. Omitting the saturation signals increased the percentage of valid diagnoses.

The results in **Table IV** show that [PI] performed better on the cardiorespiratory signals than on the electrophysiological signals. This was expected due to the relative simplicity (and robustness) of the cardiorespiratory rules as opposed to the electrophysiological rules. Generally, the rather "noisy" nature of the electrophysiological signals renders the monitoring process more complicated than that for the cardiorespiratory signals. [PI] correctly identified signal artifacts 81% of the time on the electrophysiological signals (without counting saturation) and 89% of the time on the cardiorespiratory signals. It should be noted that within the two signal categories, certain signal rules were more robust than others. The EEG rules and the EKG rules were very accurate and yielded 100% correct identifications. The EOG rules, on the other hand, were not as robust and would alert the crew with a red light whenever a signal exceeded the range displayed (instead of waiting for the signal to slowly decay as it normally happens with AC-coupled EOG's). The cardiorespiratory signals that showed the greatest number of problems were the Flow and RIP signals. The [PI] alerts about data problems for the SaO_2 and pulse-wave signals were very accurate and reflected the

many problems that the crew reported with the pulse oximeter used to record both readings.

After analyzing the signals individually, a total performance index was calculated by computing the overall percentages of correct signal artifact diagnoses. The results are tabulated in **Table V** below.

[TABLE 5 – HERE]

As the table shows, out of all the signal artifacts identified by [PI], 451 were correct diagnoses and 77 were incorrect. Without counting the 100 saturation signals for which [PI] produced a red status light, the system identified 84% of the anomalies correctly. By counting the saturation warnings, the performance decreases by about 10%. (As stated earlier, the saturation signals do not cause a problem if the astronauts are adequately trained in the recognition process and can successfully distinguish a poor signal from a simply saturated one.) The data representing correct detection of anomalies, saturation and false alarms are shown graphically in **Figures 4** and **5**.

[FIGURE 4 – HERE]

[FIGURE 5 – HERE]

Flight Performance Results From Crew Questionnaires

The crew questionnaire results are tabulated in **Table VI** below. The questionnaire was composed of yes/no questions and performance ranking questions ranging from 1 (very poor) to 5 (very good).

[TABLE 6– HERE]

A debriefing of the astronauts after the mission revealed an overall sense of satisfaction about the experiment, including the use of [PI]. In general the responses were positive, with confidence ratings also dependent on the astronauts' background and experience with physiological signal monitoring. The flight performance of [PI] and the feedback from the users also led to several modifications to improve the malfunction correction process.

DISCUSSION

Ground-Based Study Discussion

For all the data gathered, the analysis of variance (ANOVA) performed to determine the significance of several effects on the subjects' performance yielded encouraging results. The results obtained from the data analysis presented confirmed the hypothesis that a real-time expert system can positively influence subject performance in the calibration of a space life sciences experiment even with minimal training. Even though the effect of [PI] assistance on the reaction times was not statistically significant by itself ($p = 0.16$), it suggested a positive influence in improving the overall subject

performance. Aside from subjective reactions, the most evident effect of [PI] was observed in the reduction of the number of undetected anomalies, where even the influence of [PI] alone was statistically significant ($p = 0.05$) and improved by 9 anomalies out of a total of 59 presented. The number of undetected anomalies significantly decreased with the help of [PI] regardless of the day that the expert system's assistance was administered. The significance of the [PI] effect alone can be attributed to the design of the graphic user interface. Whenever an anomaly is detected, a red light next to the appropriate signal is displayed; this obviously facilitates the detection of a signal artifact, since the red light generally catches the subject's attention quite quickly.

The analysis of all the reaction times as well as the number of undetected anomalies showed that the interaction of training and [PI] assistance was also significant ($p = 0.001$ for reaction time and $p = 0.002$ for undetected anomalies). This confirms the importance of the [PI] training session on both the nature of the experiment and the use of the expert system. Even when the effect of day alone was not statistically significant ($p = 0.12$ for reaction time and $p = 0.24$ for undetected anomalies), it still suggested a positive influence. The subjects tended to perform better on the second experimental day, presumably because of the experience accumulated on the first day. We cannot, however, ignore the possibility that the re-use of the same test file may have contributed to the learning effect.

Training is required on *both* the experiment itself and the use of the expert system. There is a danger that the expert system may prove to be counterproductive if the user is not adequately trained to interpret its messages or if the familiarity with the experiment is not satisfactory. The Neurolab crew trained for several months on the

sleep experiment, including the use of the expert systems. The relatively low number of training sessions dedicated to the sleep experiment shortly before the mission, as well as the small amount of actual [PI] training within these sessions, resulted in comparable amounts and quality of [PI] training for the astronauts and the ground subjects.

In future ground-based studies it would be appropriate to increase the number of training hours until the subjects are fully confident in exercising the experimental procedures and using the expert system. It would also be advisable to use two distinct, yet statistically similar, data files for the two experimental sessions.

Poor signal quality identification was the only aspect of [PI] analyzed for this experiment. Aside from displaying data and alerting the user of a poor quality signal, [PI] shows a series of malfunction procedures on its diagnostic box: this troubleshooting capability is a very important feature that should be evaluated in future studies. An ongoing longer, multi-phased, ground-based study should provide more conclusive results on the use of expert systems not only as signal artifact identifiers, but also as troubleshooting and calibration aids.

Flight Performance Discussion

The file playback provides insight into [PI]'s performance as a monitoring system, but does not provide much information about the amount and type of interaction between the expert system and the astronauts. According to the subjects' comments, they often did not follow the malfunction procedures [PI] displayed but rather corrected the problems by remembering what they had learned in the training sessions. At this stage of the flight performance study, there is no way to determine when astronauts were

responding to the [PI] alerts and when they were responding on the basis of prior knowledge. In order to avoid this kind of uncertainty, [PI] was updated in preparation for the STS-95 mission, where the sleep experiment flew again to study the effects of microgravity on sleep and aging. The new version of [PI] requires the astronauts to click on the state light next to the poor quality signal that the system detected in order to display the corresponding malfunction procedure. During flight, [PI] recorded the astronaut mouse click times, as well as the anomaly onset times. Post-flight analysis will allow us to gather data on anomaly onset and end times.

It should also be noted that in several cases, [PI] may not have recognized a poor quality signal which a skilled operator might actually identify. Due to the nature and features of the Neurolab version of [PI], there was no reliable way of accumulating any data on the number of "false negatives"; instead, these were subjectively identified by the astronauts. This inability may be compensated for in future versions and application of [PI].

While the ground and flight results are encouraging, it would be beneficial to implement other functions that were originally part of [PI]. These include scheduling and the ability to assist astronauts with the formulation of alternate data models. These functions will require additional research in the knowledge engineering effort.

Dr. Czeisler's and Dr. West's teams are still in the process of analyzing the large amount of physiological sleep data gathered from the Neurolab mission. While at least some of the first results have been presented, at the Neurolab symposium that took place in Washington, DC in April 1999, no results have been published at the time of submission of this paper.

CONCLUSIONS

The Principal Investigator-in-a-Box expert system, designed to aid astronauts with a life sciences experiment outside their field of expertise, has been evaluated. The evaluation was divided into two parts: a preliminary ground-based study involving 12 "astronaut surrogate" subjects, and the post-flight analysis from the Neurolab Mission, in which [PI] was used to assist with the Sleep and Respiration in Microgravity Experiment.

The ground-based study revealed a positive effect of [PI] assistance on overall performance (artifact detection time and number of undetected anomalies). A cross effect of [PI] assistance and previous exposure to signal monitoring processes (training) also resulted as a significant factor in subject performance. The post-flight data analysis showed a correct diagnosis percentage of 84% of the non-saturation anomalies from in-flight. [PI]'s positive effects were supported by the positive feedback from the Neurolab astronauts. The technology should be applicable to the training and "in-flight coaching" aspects of many crew intensive experiments for the International Space Station.

ACKNOWLEDGMENTS

This research was supported by the National Space Biomedical Research Institute, NASA Cooperative Agreement NCC9-58, and the NASA Ames Research Center, grant number NCC 2-570. The [PI] Team would like to thank Dr. Charles Czeisler, Dr. Derk-Jan Dijk, Dr. James Wyatt, Eymard Riel, Joe Rhonda and Karen Smith of Brigham and Women's Hospital, and Dr. John West, Dr. Kim Prisk, Dr. Ann Elliott and Janelle Fine of the Physiology/NASA Lab at the University of California, San Diego. Dr. Peter Szolovits of MIT has been a valuable guide to the expert systems field. Many thanks to the entire STS-90 Neurolab crew: Commander Scott Altman, Dr. Jay Buckey, Commander Kathryn Hire, Dr. Richard Linnehan, Dr. James Pawelczyk, Lt. Colonel Richard Searfoss, Dr. Dave Williams, Dr. Alex Dunlap and Dr. Chiaki Mukai, as well as Suzanne McCollum, Sherry Carter, Carlos Reyes and Peter Nystrom of the NASA Johnson Space Center. Thanks also to Marsha Warren for her editorial assistance and the reviewers of this paper for their valuable critiques and insights.

BIBLIOGRAPHY AND RELATED DOCUMENTS

Callini G. Assessment of an expert system for space life sciences: a preliminary ground-based evaluation of pi-in-a-box for the Neurolab sleep and respiration experiment. Master's Thesis. Cambridge, MA: Massachusetts Institute of Technology; Sept. 1998

Colombano SP, Statler IC, Frainer RJ. The astronaut science advisor: Trade-offs between communications with the ground and 'onboard intelligence'. Proceedings of Computing in Aerospace 9. San Diego, CA; 1993.

Engle J, Bogart D, Marinuzzi J. Prelaunch expert system for space shuttle propulsion system health. AIAA, SAE, ASME, and ASEE, Joint Propulsion Conference, 26th. Orlando, FL; Jul. 16-18 1990.

Groleau N. Model-Based Scientific Discovery: A study in space bioengineering. Ph.D. Thesis. Cambridge, MA: Massachusetts Institute of Technology; Sept. 1992.

Feigenbaum E, Friedland P, Johnson BB et al. Knowledge-based systems in Japan. Report for the JTEC panel, Loyola College. Baltimore, MD; May 1993.

Friedman CP, Wyatt JC. Evaluation methods in medical informatics. New York, NY: Springer Inc; 1997.

Glass BJ. Thermal expert system (TEXSYS): Systems autonomy demonstration project. NASA Technical Report NASA-TM-102877; Oct. 1992.

Guo T. An SSME high pressure oxidizer turbopump diagnostic system using G2(tm) real-time expert system. Third Annual Health Monitoring Conference for Space Propulsion Systems. Cincinnati, OH; Nov. 13-14 1991.

Hazelton LR; Groleau N; Franier RJ et al. PI in the sky: The astronaut science advisor on SLS-2. Proceedings of the Sixth Annual Space Operations and Research Conference. Houston, TX; Aug. 1993.

Koons HC, Gorney DJ. Spacecraft environmental anomalies expert system. Aerospace Corp. Technical Report, AEROSPACE-ATR-88(9562)-1: Dec. 1988.

Lafuse SA. Development of an expert system for analysis of shuttle atmospheric revitalization and pressure control subsystem anomalies. SAE, International Conference on Environmental Systems, 21st. San Francisco, CA; July 15-18 1991.

Morris K. Expert system solutions to space shuttle payload integration design automation problems. AIAA, Aerospace Sciences Meeting, 27th. Reno, NV; Jan. 9-12 1989.

Prerau DS. Developing and managing expert systems: Proven techniques for business and industry. Reading, MA: Addison-Wesley Publishing Company; 1990.

Schwuttke UM, Quan AG, Angelino R, et al. Marvel: A distributed real-time monitoring and analysis application. Innovative Applications of Artificial Intelligence 4. Proceedings of the IAAI-92 Conference, AAAI Press/MIT Press; 1992.

Shankar D. Shuttle AI applications. AIAA, Aerospace Sciences Meeting, 27th. Reno, NV; Jan. 9-12, 1989.

Smith RL. Fault tree analysis and diagnostics development for pi-in-a-box with the neurolab sleep and respiration experiment. Master's Thesis. Cambridge, MA: Massachusetts Institute of Technology, June1997.

Young LR, Colombano SP, Haymann-Haber G et al. An expert system to advise astronauts during experiments. Proceedings of the International Astronautical Congress. Malaga, Spain; 1989.

REFERENCES

1. Franier RJ, Groleau N, Hazelton LR et al. PI-in-a-Box: A knowledge-based system for space science experimentation. AI Magazine 1994; 15(1):39-5.

2. Gundel A, Polyakov VV, Zulley J. The alteration of human sleep and circadianrhythms during spaceflight. J Sleep Res 1997; 6:1-8.

3. Kao CY. Automated Spacelab stowage expert system for SLS missions. I-SAIRAS '90. Proceedings of the International Symposium on Artificial Intelligence, Robotics and Automation in Space. Kobe, Japan; Nov. 18-20 1990:167-170.

4. Lauriente M, Rolincik M, Koons HC, Gorney D. An on-line expert system for diagnosing environmentally induced spacecraft anomalies using CLIPS. The Sixth Annual Workshop on Space Operations Applications and Research (SOAR). 1992:329-339.

5. Polyakov VV, Posokhov SI, Ponomaryova IP, et al. Sleep in spaceflight. Aerosp Med 1994; 28:4-7.

6. Szolovits P. Uncertainty and decision in medical informatics. Methods Inf Med 1995 34(1-2): 111.

7. Wang L, Fletcher M. A real-time navigation monitoring expert system for the Space Shuttle Mission Control Center. SpaceOps 1992. Proceedings of the

Second International Symposium on Ground Data Systems for Space Mission Operations; 1992:591-599.

8. **Young L.R. PI-in-a-Box. Journal of the Society of Instrument and Control Engineers, 1994; 33(2):119-22.**